Assignment 8.3 Step 1 of Final Project

Basak Atanu

05-14-2022

Introduction

I want to do exploratory data analysis on loan defaulter data, This data is sourced from Kaggle, its actually financial organization data, where PII (personal identifiable information) information has been changed for the security purpose. I am looking forward to answer different questions using this data, I have mentioned those questions in the below section.

Research questions

1. What attributes affect loan default and what are some major reasons behind it?

2. Is there any co-realation between different attributes of loan default data and gereral loan data?

3. I think, Income having a direct effect on loan default, because low income could cause default for loan payment. is it true?

4. Can I predict if the loan will go to default if I have employment, annual salary and bank balance information?

5. Does high fico socre give lower interest retes for loan?.

Approach

First I will go through the data to understand it, then will correct the data and do transformation of the data as per requirement, I may need to create additional attributes after analyzing the data, I will be doing exploratory data analysis and do model fitting to predict the future defaulters.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

Loan Application Data from Kaggle (https://www.kaggle.com/datasets/gauravduttakiit/loan-defaulter?select=application_data.csv)

loan data from Kaggle Source (https://www.kaggle.com/datasets/itssuru/loan-data)

Loan Default Data from Kaggle (https://www.kaggle.com/datasets/kmldas/loan-default-prediction)

Required Packages

GGplot

dplyr